SYSTEMS AND METHODS FOR IDENTIFYING AND EXTRACTING DATA FROM HTML PAGES

ABSTRACT OF THE DISCLOSURE

Systems and methods for analyzing HTML formatted web pages to automatically identify and extract desired information. A computer algorithm identifies and extracts different pieces of information from different web pages automatically after minimal manual setup. The algorithm automatically analyzes pages with different content if they have the same, or similar, formats. The algorithm is fast and efficient and performs the extraction process quickly in real-time. The systems and methods are useful to build databases from unstructured web information. The algorithm can be used as an agent that captures information about products, and compares prices or other characteristics. It can also be used to populate structured databases that, given the different pieces of information, can analyze products and their characteristics. And it can also be used for data mining applications looking for patterns useful for marketing analyses, or other uses.